Batch Prompting: Efficient Inference with Large Language Model APIs

Year: 2023 | *Citations: 102* | *Authors: Zhoujun Cheng, Jungo Kasai, Tao Yu*

Abstract

Performing inference on large volumes of samples with large language models (LLMs) can be computationally and financially costly in industry and real-world use. We propose batch prompting, a simple yet effective prompting approach that enables the LLM to run inference in batches, instead of one sample at a time. Our method reduces both token and time costs while retaining downstream performance. We theoretically demonstrate that under a few-shot in-context learning setting, the inference costs decrease almost inverse linearly with the number of samples in each batch. We extensively validate the effectiveness of batch prompting on ten datasets across commonsense QA, arithmetic reasoning, and NLI/NLU: batch prompting significantly~(up to 5x with six samples in batch) reduces the LLM (Codex) inference token and time costs while achieving better or comparable performance. For state-of-the-art Chat-based LLMs, e.g., GPT-3.5 and GPT-4, we show the benefits of batch prompting also hold. Further analysis shows that the number of samples in each batch and the complexity of tasks affect its performance. Moreover, batch prompting can be applied across different reasoning methods using LLMs. Our code can be found at the site https://github.com/xlang-ai/batch-prompting.